WNBA Machine Learning Predictor

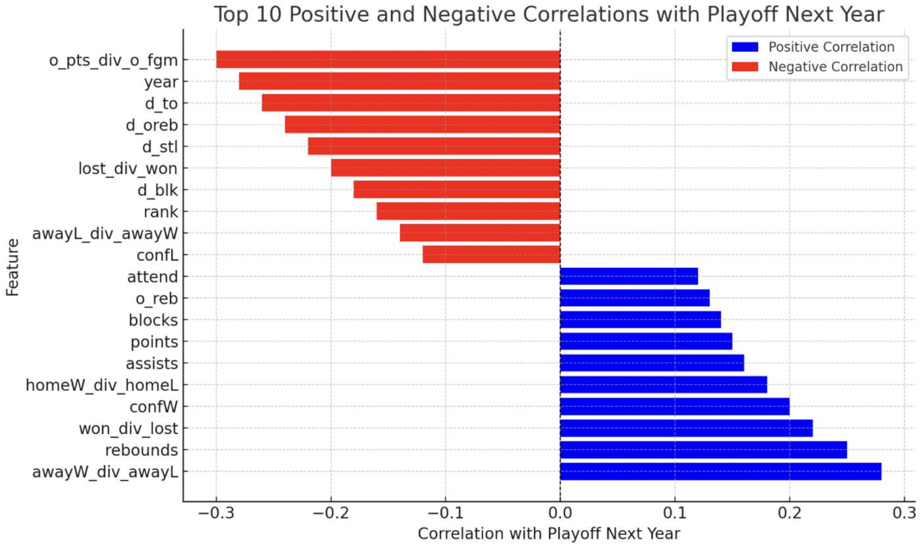
João Coelho - up202004846 - IF: 1
João Mota - up202108677 - IF: 1
Davide Teixeira - up202109860 - IF: 1

Domain Description

- The **dataset** provides comprehensive information about the **WNBA**, including **players**, **teams**, **coaches**, **awards**, and postseason **outcomes**.
- Key details span player **statistics**, team **performance metrics**, coaching **records**, and individual **awards**.
- In recent seasons, the **league** consists of 12 teams, but only 8 secure spots in the **playoffs**, emphasizing **competitiveness**.
- Insights from this data can support analyses of player **impact**, team **strategies**, and league **dynamics** over time.

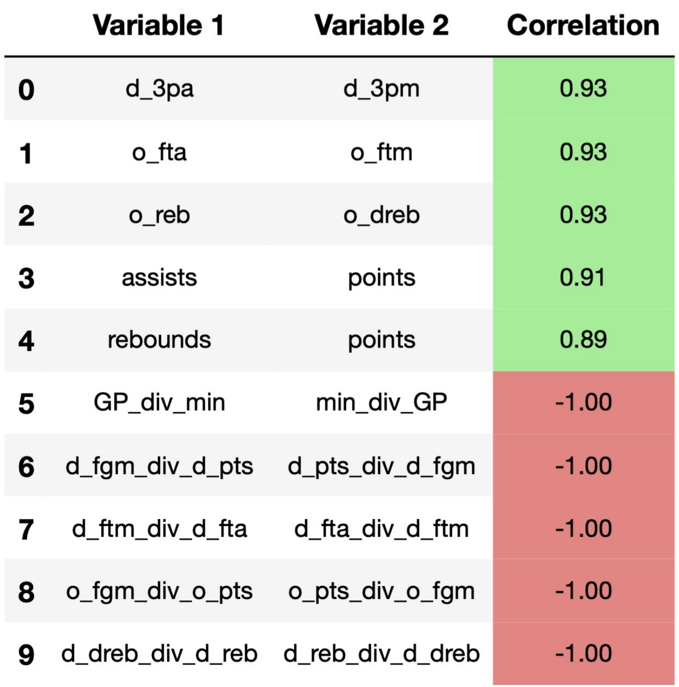
Exploratory Data Analysis

- Variables Removal:
  - NULL Values
  - lgID (Same value for all instances), tmORB, tmDRB, tmTRB, opptmORB, opptmDRB, opptmTRB, seeded, firstseason, lastseason, deathDate
- Linear Correlation with target variable
  - Feature Selection
- Conclusion
  - None of the variables is highly linearly correlated with the target variable (Playoff Next Year)
  - Importance of Multivariate Analysis



Exploratory Data Analysis (cont.)

- Linear Correlation between all variables
    - Redundant Information
        - Increases Computational Cost
    - Multicollinearity
        - Can make the model unstable and inflate the variance of coefficient estimates
        - Reduces ability to interpret feature importance
    - Risk of overfitting
        - Reduces the model's ability to generalize to unseen data



	Variable 1	Variable 2	Correlation
0	d_3pa	d_3pm	0.93
1	o_fta	o_ftm	0.93
2	o_reb	o_dreb	0.93
3	assists	points	0.91
4	rebounds	points	0.89
5	GP_div_min	min_div_GP	-1.00
6	d_fgm_div_d_pts	d_pts_div_d_fgm	-1.00
7	d_ftm_div_d_fta	d_fta_div_d_ftm	-1.00
8	o_fgm_div_o_pts	o_pts_div_o_fgm	-1.00
9	d_dreb_div_d_reb	d_reb_div_d_dreb	-1.00

Problem Definition

Creation of a Predictor that takes information about players, coaches and teams and infers if they will make the playoffs in the following year;

Prediction Task: A supervised classification problem where the target variable is whether a team makes the playoffs (1 for yes, 0 for no);

Output: List with values ranging from 0 to 1 for each team - 0 indicates missing the playoffs, 1 indicates making them;

Data Preparation

Data Cleaning:

- Decided to only use player stats, as teams are made up of players and those are what we will receive for the competition;
- Opted for awards and individual game stats, for both the regular season and postseason;

Feature Engineering:

- Assigning scores to player awards and calculating weighted scores for player performance using coefficients for key metrics;
- Aggregating these scores across players, stints, and years;
- Creating additional aggregated team-level statistics and playoff indicators;

Data Preparation

Merging Data:

- Adds all of the metrics calculated for each individual player in order to create a dataset with the teams and values representative of the talent of their players, that will be fed to the model for predictions;

Rolling Features:

- Generating rolling averages and sums for key performance metrics (e.g., scores, awards) over a three-year window to incorporate historical trends;

Experimental Setup

Using Logistic Regression, we decided to train using data from years 4 through 8 and predict the outcome of years 9 and 10;

2 different classifiers, one for each conference, and then softmax is applied to normalize the sum of each conference to 4;

Features are scaled using a Standard Scaler in order to ensure they all equally contribute to the model;

In order to simulate what is needed for the competition, we used probability prediction (proba) instead of directly predicting the labels;

Then, sort and attribute the playoff column as 1 to the 4 highest values of each conference and 0 to the remaining 2;

Results

Output is a list, to match what was required for the competition:

[1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1]

where the teams are as follows:

ATL, CHI, CON, IND, LAS, MIN, NYL, PHO, SAS, SEA, TUL, WAS

Error was 4.0

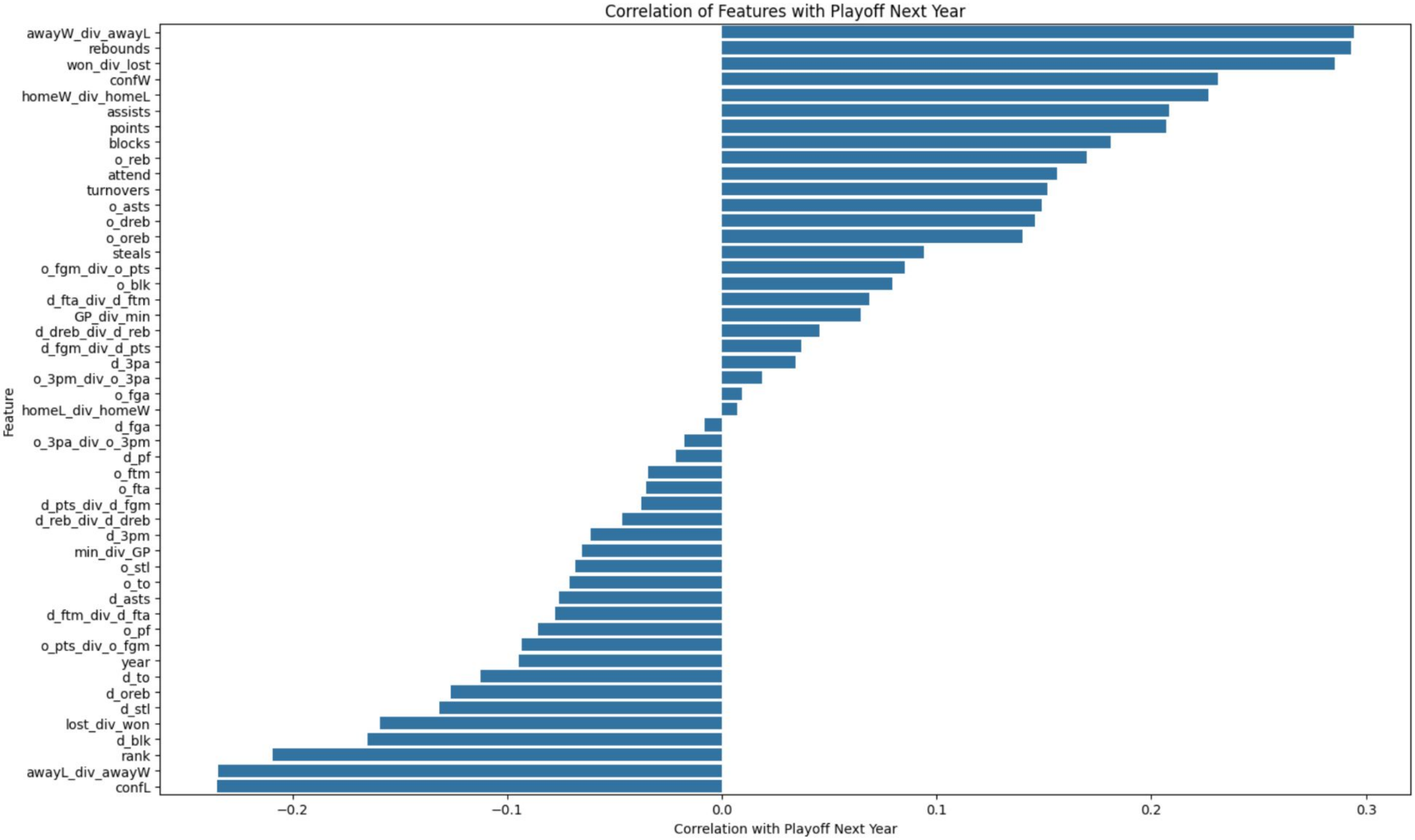
Conclusions, Limitations and Future Work

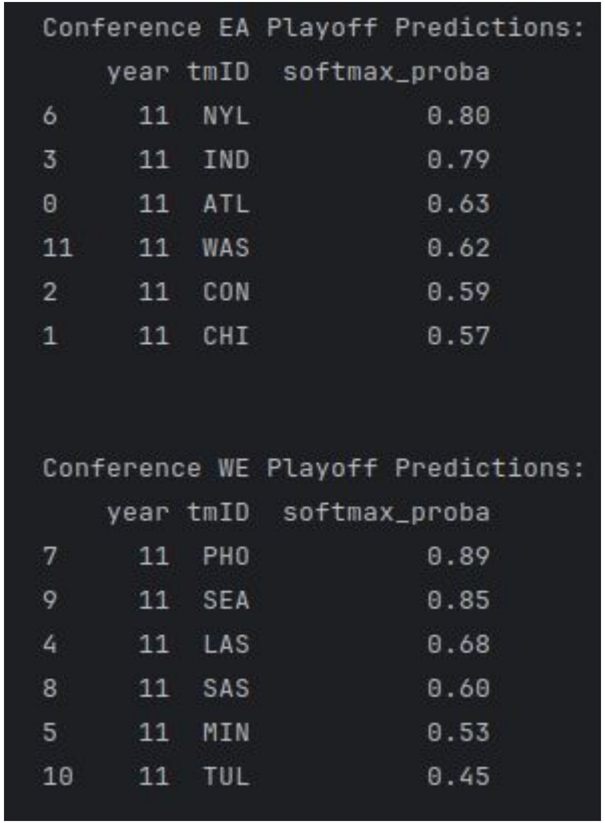
Rolling features are calculated based on a simple sum of past three years' data, without accounting for diminishing relevance of older seasons or changes in league trends;

Aggregating team and player statistics into weighted scores and rolling features allows for the model to predict without needing stats from the year it wants to predict;

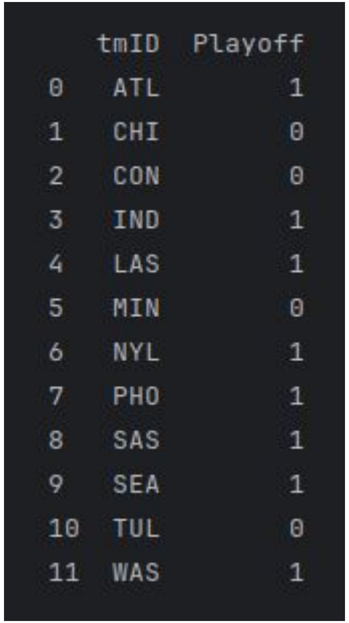
The model doesn’t account for rookies that may improve the team a lot, as well as the impact coaches may have

Annexes

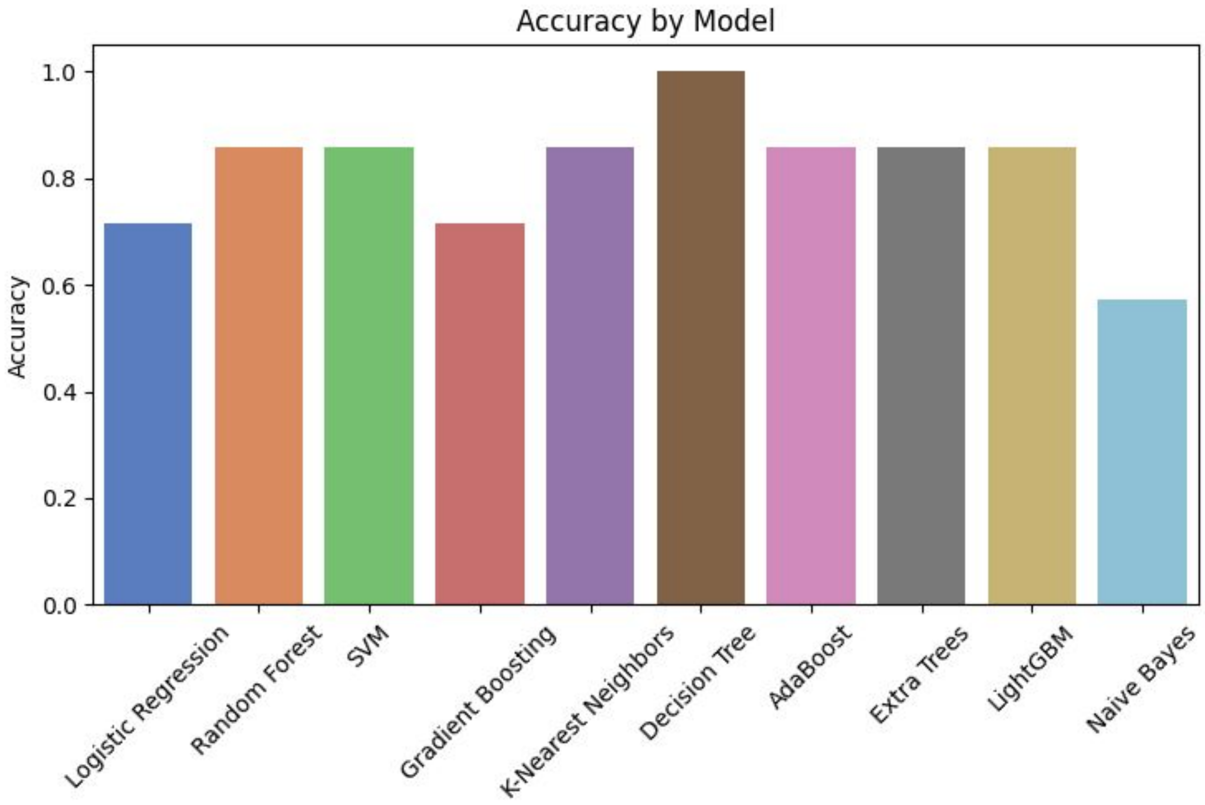

Correlation of Features with Playoff Next Year
awayW_div_awayL
rebounds
won_div_lost
confW
homeW_div_homeL
assists
points
blocks
o_reb
attend
turnovers
o_asts
o_dreb
o_oreb
steals
o_fgm_div_o_pts
o_blk
d_fta_div_d_ftm
GP_div_min
d_dreb_div_d_reb
d_fgm_div_d_pts
d_3pa
o_3pm_div_o_3pa
o_fga
homeL_div_homeW
d_fga
o_3pa_div_o_3pm
d_pf
o_ftm
o_fta
d_pts_div_d_fgm
d_reb_div_d_dreb
d_3pm
min_div_GP
o_stl
o_to
d_asts
d_ftm_div_d_fta
o_pf
o_pts_div_o_fgm
year
d_to
d_oreb
d_stl
lost_div_won
d_blk
rank
awayL_div_awayW
confL
Feature
−0.2
−0.1
0.0
0.1
0.2
0.3
Correlation with Playoff Next Year

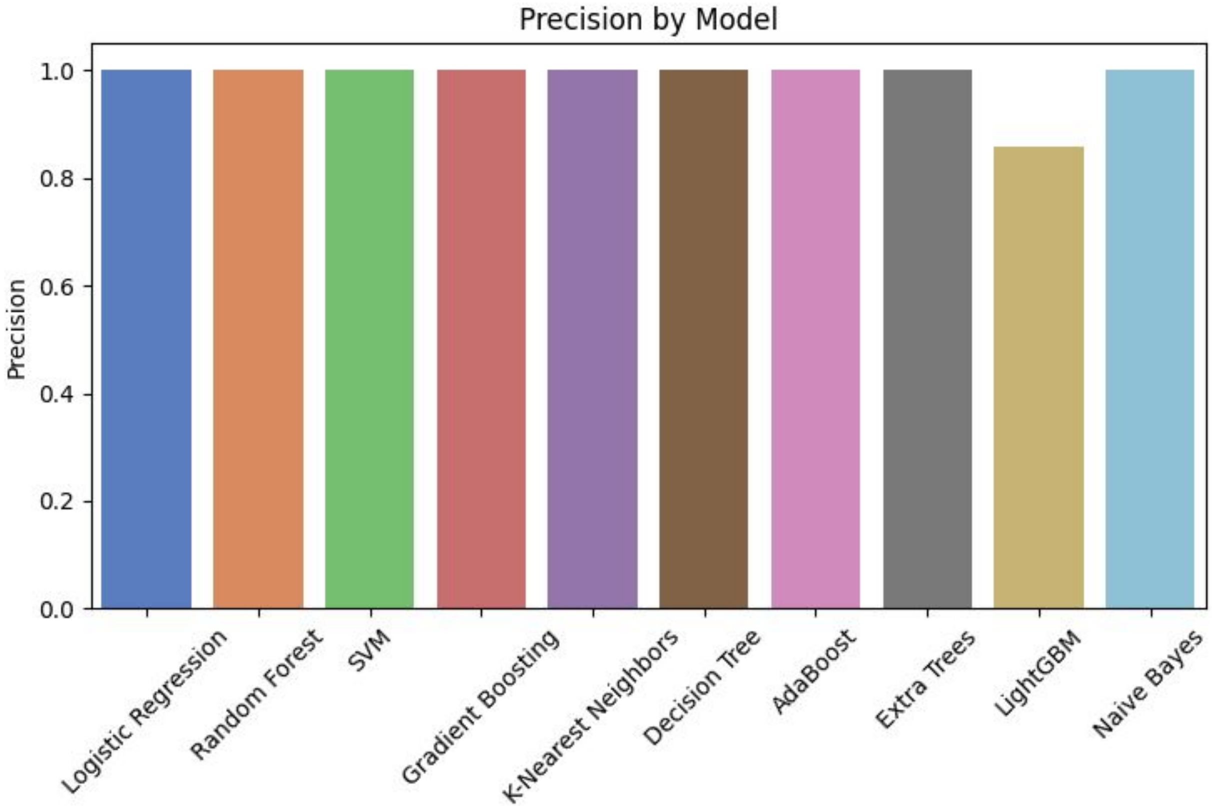


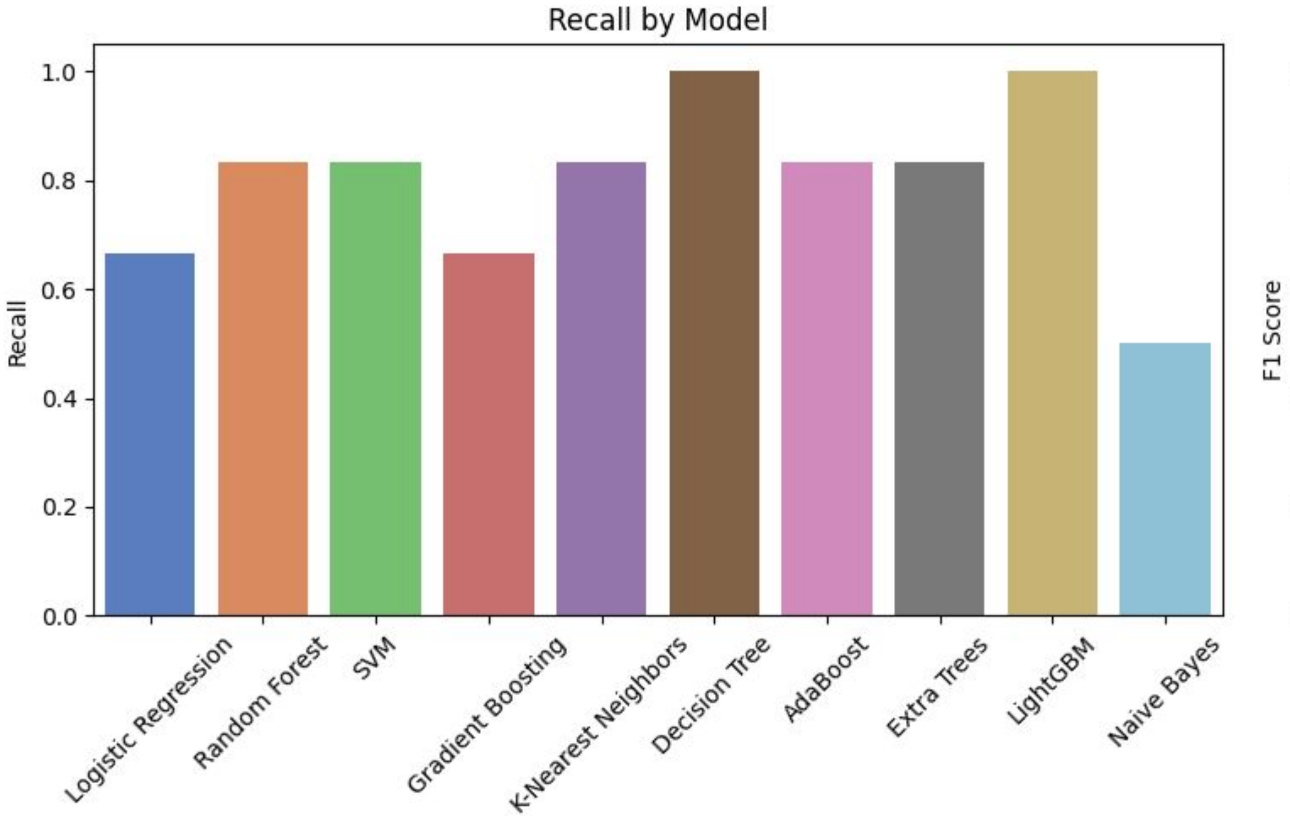
```
Conference EA Playoff Predictions:
    year tmID  softmax_proba
6     11  NYL           0.80
3     11  IND           0.79
0     11  ATL           0.63
11    11  WAS           0.62
2     11  CON           0.59
1     11  CHI           0.57


Conference WE Playoff Predictions:
    year tmID  softmax_proba
7     11  PHO           0.89
9     11  SEA           0.85
4     11  LAS           0.68
8     11  SAS           0.60
5     11  MIN           0.53
10    11  TUL           0.45
```

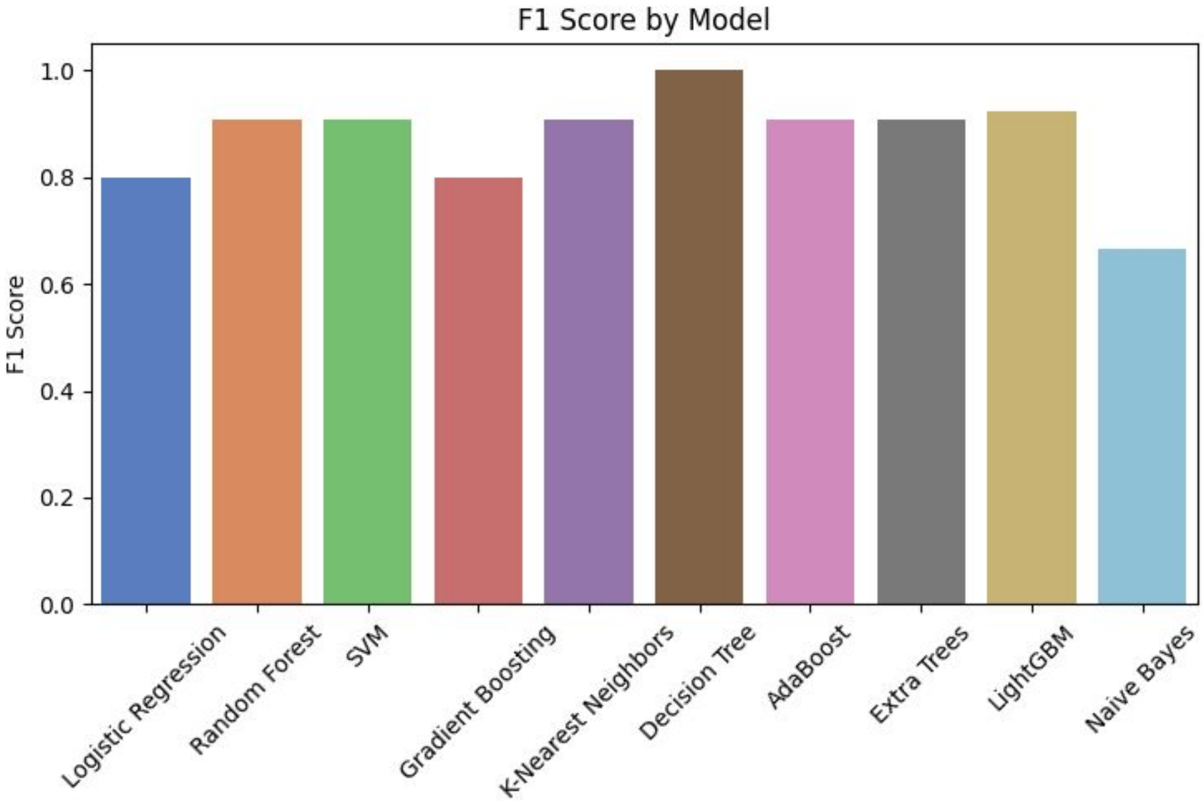


```
   tmID  Playoff
0   ATL        1
1   CHI        0
2   CON        0
3   IND        1
4   LAS        1
5   MIN        0
6   NYL        1
7   PHO        1
8   SAS        1
9   SEA        1
10  TUL        0
11  WAS        1
```

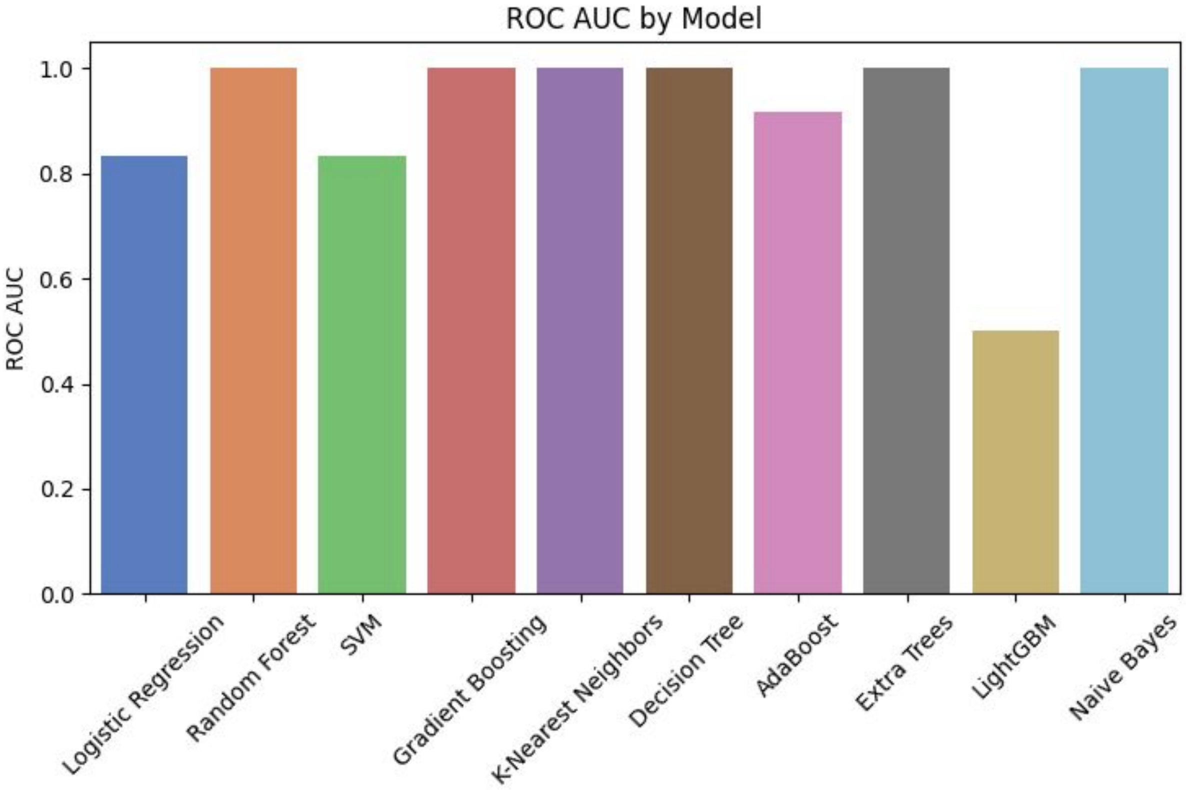

Accuracy by Model
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
Logistic Regression
Random Forest
SVM
Gradient Boosting
K-Nearest Neighbors
Decision Tree
AdaBoost
Extra Trees
LightGBM
Naive Bayes

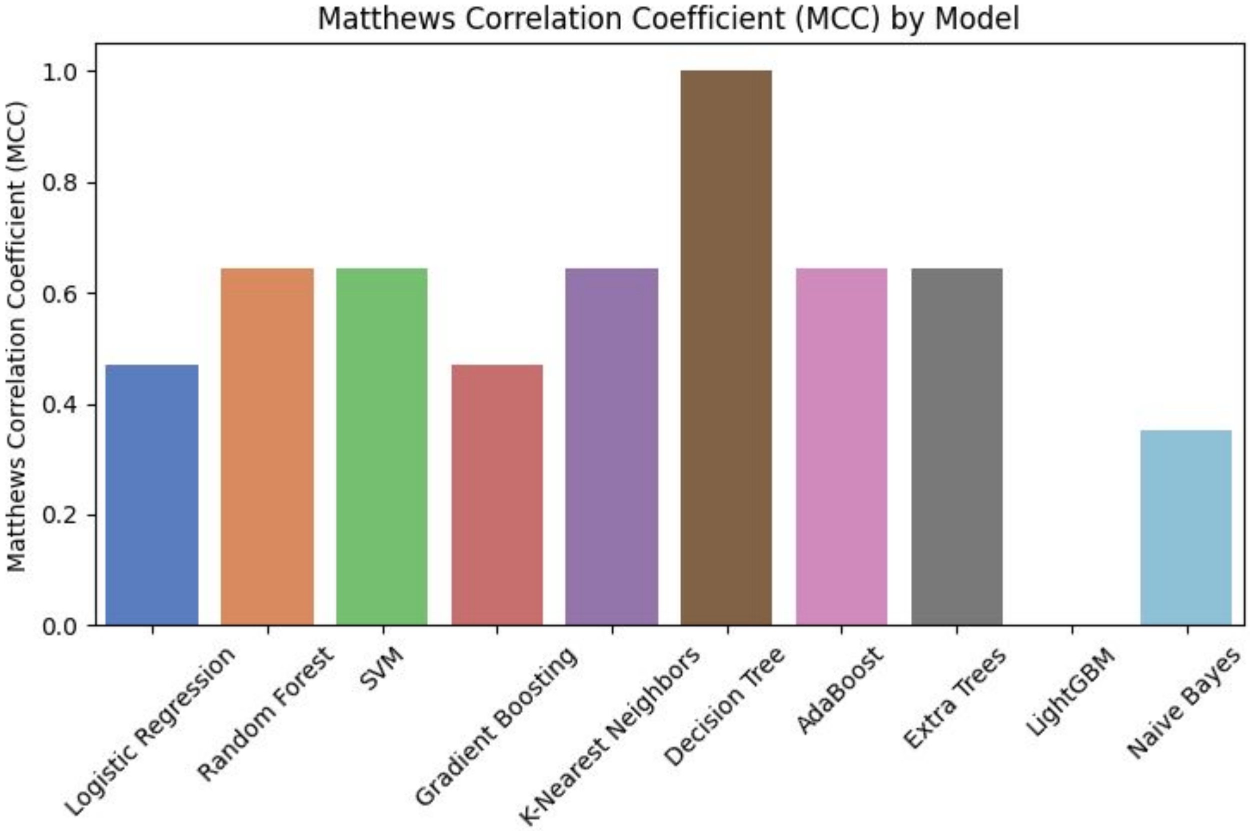

Precision by Model
Precision
1.0
0.8
0.6
0.4
0.2
0.0
Logistic Regression
Random Forest
SVM
Gradient Boosting
K-Nearest Neighbors
Decision Tree
AdaBoost
Extra Trees
LightGBM
Naive Bayes

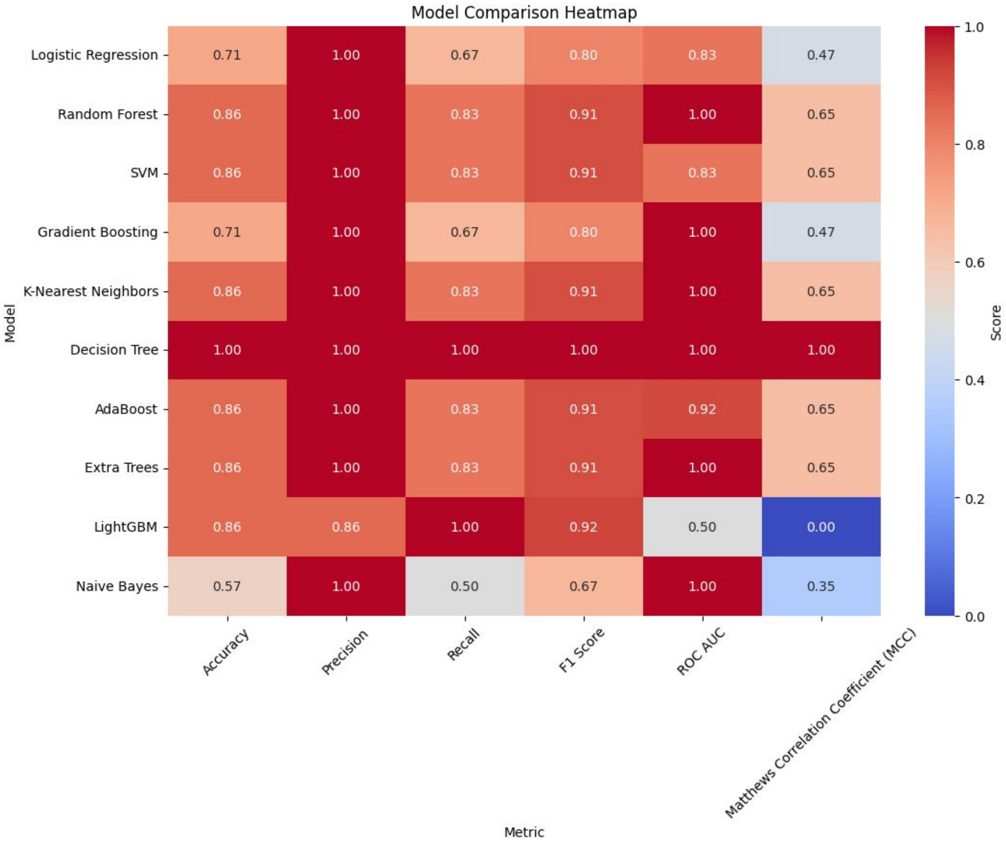

Recall by Model
Recall
1.0
0.8
0.6
0.4
0.2
0.0
Logistic Regression
Random Forest
SVM
Gradient Boosting
K-Nearest Neighbors
Decision Tree
AdaBoost
Extra Trees
LightGBM
Naive Bayes
F1 Score



F1 Score by Model
F1 Score
1.0
0.8
0.6
0.4
0.2
0.0
Logistic Regression
Random Forest
SVM
Gradient Boosting
K-Nearest Neighbors
Decision Tree
AdaBoost
Extra Trees
LightGBM
Naive Bayes


ROC AUC by Model
ROC AUC
1.0
0.8
0.6
0.4
0.2
0.0
Logistic Regression
Random Forest
SVM
Gradient Boosting
K-Nearest Neighbors
Decision Tree
AdaBoost
Extra Trees
LightGBM
Naive Bayes


Matthews Correlation Coefficient (MCC) by Model
Matthews Correlation Coefficient (MCC)
1.0
0.8
0.6
0.4
0.2
0.0
Logistic Regression
Random Forest
SVM
Gradient Boosting
K-Nearest Neighbors
Decision Tree
AdaBoost
Extra Trees
LightGBM
Naive Bayes


Model Comparison Heatmap
Model
Metric
Score
Accuracy
Precision
Recall
F1 Score
ROC AUC
Matthews Correlation Coefficient (MCC)
Logistic Regression 0.71 1.00 0.67 0.80 0.83 0.47
Random Forest 0.86 1.00 0.83 0.91 1.00 0.65
SVM 0.86 1.00 0.83 0.91 0.83 0.65
Gradient Boosting 0.71 1.00 0.67 0.80 1.00 0.47
K-Nearest Neighbors 0.86 1.00 0.83 0.91 1.00 0.65
Decision Tree 1.00 1.00 1.00 1.00 1.00 1.00
AdaBoost 0.86 1.00 0.83 0.91 0.92 0.65
Extra Trees 0.86 1.00 0.83 0.91 1.00 0.65
LightGBM 0.86 0.86 1.00 0.92 0.50 0.00
Naive Bayes 0.57 1.00 0.50 0.67 1.00 0.35
1.0
0.8
0.6
0.4
0.2
0.0